ABSTRACT

A system and method facilitating speech detection and/or enhancement utilizing audio/video fusion is provided. The present invention fuses audio and video in a probabilistic generative model that implements cross-model, self-supervised learning, enabling rapid adaptation to audio visual data. The system can learn to detect and enhance speech in noise given only a short (*e.g.*, 30 second) sequence of audio-visual data. In addition, it automatically learns to track the lips as they move around in the video.